DOCUMENT RESUME

ED 247 240 TM 840 385

AUTHOR Gonzalez-Tamayo, Eulogio
TITLE An Item Content Criterion of Debiasing a Test.
PUB DATE 84
NOTE 29p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Difficulty Level; *Hispanic Americans; Item Analysis; Language Dominance; Minority Groups; *Test Bias; *Test Items; Test Validity
IDENTIFIERS California Achievement Tests; *Item Content Criterion

ABSTRACT

An item content criterion, independent of the test psychometric characteristics, of classifying items as biased is described. It was used with a sample of female adults in a training program for administrative secretaries. The minority group in the study were Hispanic immigrants. The majority group was a mixture of Blacks, English speaking Caribbeans, and Hispanics born in the continental United States or those who attended American schools. The impact of removing biased items on the validity coefficient and the groups' mean scores was investigated. The elimination of biased items reduced the mean difference in the predictor between the majority and minority groups, thus improving the test's validity. The most discrepant items by the Transformed Item Difficulty (TID) method and the items with unacceptable difficulty differences between the groups were compared with the items detected as biased by the item content criterion. The greatest similarity was found between the item content criterion and the unacceptable difficulty differences.
(Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

4/25/84

ED247240

AN ITEM CONTENT CRITERION OF DEBIASING A TEST

Eulogio Gonzalez-Tamayo

New York, 1984

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Eulogio Gonzales-Tamayo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 840 385

SUMMARY

An item content criterion, independent of the test psychometric characteristics, of classifying items as biased is described and used in a sample of adults involved in a training program. The elimination of biased items reduced the mean difference in the predictor between the majority and the minority groups improving the validity of the test. The most discrepant items by the Transformed Item Difficulty (TID) method and the items with unacceptable difficulty differences between the groups were compared with the items detected as biased by the item content criterion. The greatest similarity was found between the item content criterion and the unacceptable difficulty differences.

INTRODUCTION

The concepts of fairness and bias have been part of the discussion about the aptitude testing controversy and the issues raised by them have acquired some prominence in the scientific journals. Although the concepts appear interchangeable at times, they are not; fairness is related to the uses of a test while bias deals with the structure of the test regardless of its use. A test has to be unbiased before we can talk about its fair use, but the use of an unbiased test can still be unfair to a particular group.

Fairness arises when the differences between two groups on the predictor (test) are larger than the differences on the performance criterion. Similar differences both in the predictor and in the criterion would indicate that the test reflects differences in performance and there is no reason for talking about possible unfairness of the test against one group.

Bias relates to the test itself regardless of how it is used. Groups score differently in aptitude tests; if these differences correspond to actual differences in the ability the test is intended to measure, the test is unbiased regardless of its use; but if the differences in the test reflect entirely or partially some other variables where the two groups differ markedly, the test is biased against the lower scoring group. The choice of a standard (test) and the items by which a given function is to be judged, is the most likely source of bias (Gordon & Terrell, 1981).

Factor analysis has been used as a technique to find biased items. The technique describes the interrelationships among variables (test items) in terms of a smaller number of dimensions (factors). If the found dimensions are different for two groups, then the test is biased; if the factors are similar, the test represents the same dimensions in both groups and in the same way, the test, therefore, is not biased. Factor analysis studies have shown very similar factor structures across groups both in terms of the strength of the first factor

and in the factor patterns (Humphreys & Taber, 1973). The in-put in factor analysis is the correlation coefficients amomg items and the outcome is a structure of performance of two groups. Differences between groups are not of kind but rather of degree (Humphreys, 1973). Two groups probably place different emphasis on different manifestations of a particular skill; but it is very unlikely that they only use one of several manifestations and not the others. The structure is likely to be the same in both groups (no differences in factor analysis) but members of one group will probably be more familiar with some manifestations than the members of the other group (different test performance and possible bias).

Predictive validity, including the validity coefficient and the parameters of the regression equation, has also been used as a technique to study possible bias in the prediction rather than in the test itself (Cleary, 1968; Jensen, 1980; Anastasi, 1982; Wigdor & Garner, 1982). The results have been very similar validity coefficients and regression parameters for both majority and minority groups with the common regression line overpredicting the minority mean criterion score. Differential validity and meta-analysis of differential validity have very little power to detect important types of bias, and differential prediction analyses have greater statistical power but are subject to problems (Drasgow, 1982).

Most of the studies and techniques of detecting bias have been through the analysis of the characteristics of the items. Test items represent events or instances in which an ability can be shown, and it is very likely that cultural differences influence on the way an ability is manifested. Two groups may have the same ability but the events included in the test are more familiar to one group than to the other. Test developers, regardless of how hard they may try, do not produce items is a vacuum but rather within the field of their experience. Most studies of item analysis have shown at least some, and occosionally many, items with anomalous statistics among groups (Cole, 1981) and there is evidence of convergence between the different methods (Shepard & Al., 1981) indicating that bias is a valid construct in studying test scores.

Several techniques of item analysis to detect bias have been proposed. An-

goff & Ford (1973) considered items as biased if they were relatively more difficult for one group than the other; Scheuneman (1979) proposed the chi-square method comparing the actual to the expected correct item responses in both groups; Green & Draper (1972) used the point biserial correlation to investigate whether there was differential discrimination among groups; based on the latent trait theory, Lord (1980) proposed the three parameter and Wright & Stone (1979) the one parameter model. Latent trait models are the most elegant statistically and the soundest theoretically but they require a large N to really be effective in detecting bias among groups.

The picture from item analysis techniques of detecting bias is not a cohesive one; the method used makes a difference in the items detected as biased (Shepard & Al.,1981). The Transformed Item Difficulty (TID) method compounds bias with ability (Angoff, 1982). The biserial method compounds discrimination with bias; the most biased may be the best discriminating items and, when two groups differ markedly on the trait, the best discrimination is obtained when the items are completely biased, that is, they are answered correctly by all the members of the high scoring group and by none from the low scoring group. Latent trait and chi-square techniques attempt to control for ability but they control imperfectly. The distribution of ability (group dependent) has some effect on estimation of parameters and, since ability is estimated from test information, the presence of biased items affects that estimation (Ironson, 1982). Rudner (1983) demonstrated that parameter values from separate calibrations are linearly related, being invariant only in the sense that theoretically one can always recalibrate the parameters without any loss in the interpretation. Lord (1980) proposed a two stage procedure, detecting and removing the most biased items in the first stage and estimating ability from the remaining items and calculating bias using this refined measure of ability; there is no reason why the process could not continue beyond the second stage until no biased items are found in a particular stage.

DESCRIPTION OF THE PRESENT STUDY

A problem common to all statistical methods of item bias analysis, even the most sophisticated as the latent trait methods, is that they are dependent to some degree on the internal characteristics of the test. There is also no psychological rationale for deciding that a statistically biased item is indeed biased (Angoff, 1982). A criterion indepedent of the psychometric characteristics of the test, although always desirable, is in most instances difficult to obtain. We were able to devise such a criterion by analyzing the content of the items and before doing any empirical analysis of the responses. In simulation studies, the item parameters rather than the items themselves are manipulated; the evidence of validity for the detection methods would be more compelling if we could construct biased items by manipulating the content (Shepard & Al.,1981).

A second purpose was to study the impact of removing biased items on the validity coefficient and the groups mean scores. The only legitimate consideration for the use of tests in selection and admissions is their predictive validity. The test bias controversy originated when large mean differences were reported between different ethnic groups; if the removal of biased items has no important effect on group mean differences, the techniques would be a theoretical exercise with no practical implications. This impact, however, has not been widely analyzed. Statistical procedures have become an end in itself where results are carefully tabulated and reported and nothing further is done (Scheuneman, 1982); an outstanding issue of item bias techniques is the effect of removing the biased items, whether or not the removal reduces the adverse impact on groups without impairing the predictive validity (Ironson, 1982).

Sample

A group of adults participating in a training program for administrative

secretaries in New York City provided the opportunity to define and apply an item content criterion of detecting biased items. The number of students, during a three year period from 1979 to 1982, was 149. About 75% of them were of Hispanic ethnicity and the rest belonged to other minorities. All students were women, all came from low income families, and about 30% were receiving public assistance. The students either had a high school diploma (78%) or were in the process of obtaining their GED. The age range was between 18 and 45 years with a median of about 23 years of age.

The minority group (81 subjects), to be defined later, were all Hispanic immigrants. The majority group (69 subjects) was a mixture of blacks, English speaking Caribbeans, and Hispanics (46%) either born in the continental U.S. or immigrants who attended American schools for six years or more.

Group Classification

Language dominance was the criterion to separate the students into two clearly identifiable groups: those who had Spanish and those who had English as their primary language of communication as determined by the amount of schooling in English institutions. Primary rather than first language was considered as more appropriate because most of the students had Spanish as their first language, but many of them functioned better in English because of their education.

If the students had received six or more years of their education before graduation from high school in English institutions, they were considered as English dominant or the majority group; if the students had attended English schools for less than six years, they were considered as Spanish dominant or minority group.

The number six was arbitrary but it was based on the following considerations. First, the greater the number of years in English schools, the better the students would be able to function in English. Second, the younger they were when they came to the USA, the easier it would be for them to learn the new language.

Third, based on our experience with the students, we realized that those who had not been in grade English schools were not English dominant either because they were relatively too old or had too much education in Spanish when they started learning English. The six years of English grade schooling made the students relatively young (about 12 years of age) when they started learning English and it also gave them a relatively large amount of exposure to English (half the number of years of their education). A final consideration in selecting this criterion was that the information was available from the students' records and no alternative was completely satisfactory or better than the one selected.

Item Content Criterion of Item Bias

The instrument analyzed for possible bias in this particular situation was the vocabulary sub-test of the California Achievement Test (CAT, 1977). The criterion for deciding whether the item was biased against Hispanics, the minority group, was the existence or non-existence in the Spanish language of the word representing the item. If the word is used both in Spanish and English, even though the pronunciation and spelling may be different in both languages, the item was considered as fair for both groups (vigor-vigor, disadvantage-desventaja). If the word appears only in the English language (loath), it was considered as biased against the minority group. All the students accepted into the program had a fair understanding of English both in reading comprehension, at least 7.0 GE in the CAT, and in verbal communication as determined by the interview.

There was the assumption that both groups had had similar opportunity of learning the word that appeared in both languages, while the words that appeared only in English would be more related to the majority than the minority experience. The criterion was considered reasonable because the purpose of the test in the program was to select candidates who would do well in training and on the job; it would not have been a good criterion if the purpose were to predict achievement

in English; but success in the program was more than familiarity with few English words. Jensen (1980) considered that if the test involves the native language of the major group and the minor group has a different language, the test will most likely be positive biased in favor of the major group.

The vocabulary sub-test used in the study is an achievement test measuring understanding or knowledge of words. The test measures a developed mental ability (synonyms, antonyms) as it is manifested in the events included in the test. An ability can be measured in a meaningful way only if the person has been exposed to the word. The test was developed for English speakers and it, therefore, was not intended to measure language ability. Words not used in Spanish would give an advantage to the majority group in showing this ability. Only by selecting words used in both languages would the test be fair for both groups, because only then would it measure developed ability rather than language dominance. Vocabulary has been recognized not only as an excellent measure of verbal comprehension (Sternberg & Powell, 1983) but also as one of the best single indicators of a person's overall level of intelligence (Jensen, 1980; Matarazzo, 1972).

This kind of criterion of identifying biased items against Hispanics could be relevant in aptitude and ability sub-tests which include synonyms, antonyms, verbal analogies, and even verbal reasoning. Familiarity or exposure to the words is a prerequisite to ascertain similarities, differences, or relationships between words.

The same reasoning can be applied to any aptitude test given to two groups with different experiences; a developed ability can only be measured when there has been an opportunity to show it and the ability can only be measured in a fair way by includings events in the test where both groups had similar opportunity of showing the ability. Every test tends to favor the culture in which it was developed (Anastasi, 1982). The problem in most cases, as opposed to the present study, is that there is no way of assessing the similarity of opportunities until after the test has been used and, then, only in a very undefined way through the analysis of anomalous behavior of some items in relation to other items.

Instruments

The predictor was the California Achievement Test (CAT) which all the students are required to take six weeks before leaving the program; the length of the program was 30 weeks. The pre-test scores were not used because many students were tested outside the program and we had the total score but not the answers to the individual items. The vocabulary sub-test contains 30 items dealing with the meaning (synonyms and antonyms) of single words.

The criterion of performance was an equally weighted combination of progress in English, stenography, typing, mathematics, effort, and success in placement. Each component was rated from zero to three based on the following predetermined criteria of progress and success.

The monthly teachers' scores (one to four) in English, mathematics, and effort, were averaged for the year (seven months). The average scores were translated into the new scale in the following way: an average score of 3.5 or more was 3 in the new scale, an average of between 3.0 and 3.4 was 2, and average of between 2.5 and 2,9 was 1, and an average of less than 2.5 was zero. The conversion into the new scale was necessary because none of the students were rated, as an average, below 2.0 and using the original scores would overweigh these components in relation to the others for which there were not teachers' evaluations.

In stenography, if the student at the end of the program was doing transcriptions at 80 words per minute or more it was rated 3 in the new scale, if she was doing 70 wpm was rated 2, 60 wpm was rated 1, and less than 60 wpm was zero.

In typing, if the student was doing 50 wpm net or more she was rated 3, between 40 and 49 wpm was rated 2, between 30 and 39 was rated 1, and less than 30 wpm was rated zero.

Success in placement was rated 3, if the student was placed as secretary; it was rated 2, if she was placed as clerk typist before the end of the program

(easy placement); it was 1, if she was placed as clerk typist after the end of the program (difficult placement); then zero, if she was not placed.

All the students records were analyzed and translated into the new scale. The total was a simple aggregation of the scores on each of the six individual components with the lowest possible score being zero and the highest possible score 18 points.

RESULTS OF THE STUDY

The scores on the vocabulary sub-test and the total ratings on the criterion of performance in the training program were separately analyzed for the majority and the minority groups.

Analysis of Groups Means

Based on the item content criterion of bias, 8 items out of a total of 30, or 27% of the items, were classified as biased in favor of the majority group. Table 1 contains the information about the means and standard deviations for both

---

Table 1 about here

---

the majority and the minority groups on the criterion of performance, on the total test, on the debiased test, and on the biased items.

Both groups scored very similar on the criterion and the differences between their means was not significant at the .05 level. The majority group, however, scored higher than the minority group.

No significant mean group differences were expected in the predictor, if the test was working properly, because no differences were found in the criterion. Significant differences in the predictor would indicate that the test was favoring one group, if used in selection, when no differences in performance were found between the two groups.

There were significant differences between the two groups, favoring the majority group, in the mean score on the total test and on the biased items, but no significant differences were found in the debiased test. The items previously classified as biased were clearly responsible for the differences in the total test, an apparent indication of probable bias. In the debiased test, the majority

group scored higher than the minority group, as expected from their scores on the criterion, but the differences, as in the criterion, were not significant.

The average group differences in difficulty on the total test, as reported in table 2, were about eight percentage points (83% to 75% for the majority and

---

Table 2 about here

---

and the minority group respectively). The sub-test was easy for both groups. In the biased items, the differences in difficulty were much greater, about 26 percentage points, in favor of the majority group. There is some strong indication that the items classified as biased by the item content criterion, are indeed biased according to the average performance of the two groups in the total test and in the biased items; there is some kind of anormality in the biased items in relation to the total test.

There were no significant group differences, as an average, in the debiased test; the difference was only two percentage points in favor of the majority group. In the present study, the biased items were responsible for the group differences in the total test. Eliminating items detected as biased by an item content criterion makes the test easier for both groups, and the mean group difference is also substantially reduced.

Validity Coefficients

The product moment correlation coefficients between the scores on each of the predictors (total test, debiased test, and biased items) and the scores on the criterion of performance were calculated and they are reported in table 3. The

---

Table 3 about here

---

correlation between both parts of the predictor (unbiased and biased items) was

.33 for the majority group, .48 for the minority, and .39 for the total of both groups. The two parts were modestly related although in a significant way. The coefficients were lower than usually found in studies correlating similar tests (both parts are vocabulary tests), indicating that both parts were not measuring the same thing to a substantial degree.

The validity of the debiased test (correlation between predictor and criterion of performance) was higher than the validity of the total test in both the majority (.41 to .38) and the minority (.51 to .42) groups. There is no technical reason to include the biased items into the prediction because they do not improve the prediction; as a matter of fact, the test is a better predictor without than with the biased items. The lack of improvement in prediction is the most powerful argument about the biased nature of the items.

The correlation between the entire test and the criterion was lower than the correlation of one of its parts (the debiased test) both in the majority and in the minority groups. The finding seems to be at odds with the theoretical procedure for the multiple correlation. In the study, all the theoretical assumptions of the multiple correlation were not met; this, however, does not explain that the total was lower than the partial correlation. A possible explanation is that the biased items were not working in the same way as the unbiased items and when both were combined a new relationship appeared that was not identical to the previous one.

ITEM CONTENT CRITERION AND TWO OTHER METHODS OF IDENTIFYING BIASED ITEMS

The item content criterion of classifying items as biased, used in the present study, is based on the nature of the items regardless of their statistical characteristics. All the published methods of detecting bias in the items are based on the psychometric characteristics of the test. We did a comparison of two numerical methods of detecting biased items with the item content criterion.

One of the methods was the Transformed Item Difficulty (TID) method proposed by Angoff & Ford (1973). The difficulty of the items was transformed to z-scores based on the normal distribution and then converted to delta-values through a simple linear transformation. The purpose of this transformation is to convert the percent incorrect (unequal units) into an interval scale (z-scores) and to make all the scores positive (delta values). These delta values are represented by dots in figure 1. The major axes of the elipse for the total test, for the test debiased by the item content criterion, and by the TID method are

---

Figure 1 about here

---

also represented and they were found by the formulas proposed by Angoff (1982). The 8 items with the largest distances to the axis of the elipse for the total test were selected as biased and compared with the 8 items classified as biased by the item content criterion.

The other numerical method was a simple percentage point procedure based on the straight differences in difficulty between the two groups. Three cases were analyzed (10, 15, and 20 percentage points difference) because there was no prior indication of how large the differences had to be in order to be considered as anomalous or discrepant.

Items Detected as Biased by the TID Method

Table 4 reports the origin of the most discrepant items by the TID method

---

Table 4 about here

---

whether they are from the biased or the unbiased part of the test according to the content criterion.

There is no agreement between the two methods. Only 2 of the 8 most discrepant items by the TID method are from the group of 8 items classified as biased by the item content criterion, and these two items are the ones with the smallest groups difference in difficulty among the 8 items. Five of the items have positive distance difference indicating that they are biased againsnt the majority group. When the number of most discrepant items by the TID method is increased to 18, only 5 of the 8 items classified as biased by the item content criterion are included in the group. It is obvious that the two methods are detecting items as biased for different reasons.

Eliminating the most discrepant items does not substantially change the difficulty of the test for either group nor does it reduce the mean group difference in diffiulty as reported in table 5. The mean diffiulty in the total test

---

Table 5 about here

---

was 83% for the majority and 75% for the minority group. Eliminating the 8, the 13, and the 18 most discrepant items makes a negligible change in the mean difficulty for both the majority (83% to 84%) and the minority group (75% to 77%) with the same difference between the means difficulty (8 or 7 percentage points). The TID method does not seem to change the relative standing of the two groups on the trait and although of some theoretical interest does not seem to have any practical consequence in the controversy about tet bias at least in the present sample.

Unacceptable Difficulty Method

Nungester (1977) reported that the removal of items with unacceptable discrimination or difficulty indices improved the reliability and validity of the tests to a greater extent than the removal of items detected as biased by three psychometric methods which showed very small improvements. In our example, the removal of items detected as biased by the item content criterion improved the validity of the test for both the majority and the minority groups; it was natural to investigate how the unacceptable difficulty difference and the item content methods do in relation to each other. In table 6 the results are presented of the origin of the items with unacceptable difficulty (defined as 10, 15, and 20 percentage points difference in difficulty between the two groups) in relation

---

Table 6 about here

---

to the items classified as biased or unbiased by the item content criterion.

This simple method of straight percentage difference in difficulty provides some idea of how large the group differences have to be in order to be considered as discrepant and the items as biased. Group differences of 10 percentage points in item difficulty can be found even in items classified as unbiased by the item content procedure. If group differences are 15 percentage points or larger, then, all the items selected as biased, with the exception of one, were indeed classified as biased by the item content criterion, and none of the items classified as unbiased, with the exception of one, were selected by the unacceptable difficulty difference method. It seems that a good rule of thumb of detecting bias in the present study would be group differences in difficulty between 10 and 15 percentage points.

The average group difference in difficulty in the total test was 7.97 percentage points. The standard deviation for the minority group converted into percentage, was 10.73 percentage points. A good approximation to the rule of thumb

would be the average group difference (7.97) plus one half the standard deviation (5.37) for the minority group or 13.34 percentage points group difference in difficulty. Applied to the present study, this criterion produced the same results as the 15 and 20 percentage points method: it identified 7 out of 8 biased items and it selected as biased only one out of 22 items classified as unbiased by the item content criterion. The method made 88% positive correct decisions (eliminated items classified as biased by the item content criterion) and only 5% incorrect negative decisions (eliminated items that were not classified as biased); the total number of correct decisions (positive and negative) was 93% of all the items. The method makes concessions for group differences in ability since discrepancy is defined as departure from the average group difference in difficulty rather than simple differences in difficulty.

CONCLUSION

A substantial number of items in the vocabulary sub-test of the CAT were classified as biased against a Spanish dominant group according to an item content criterion (the word exists in English but not in Spanish).

The results of the study indicate that it is possible to reduce the mean group differences on the predictor by eliminating items that are biased without impairing the predictive validity of the test. The elimination of biased items produced both a substantial reduction in mean difference between the majority and the minority groups and an increase in the validity of the test.

A simple method of detecting biased items (mean group difference in difficulty plus one half of one standard deviation in percentages) worked very well in detecting the items classified as biased by the item content criterion.

The TID method did not work very well in the present study; it did not detect a substatial number of items classified as biased by the item content criterion and eliminating the biased items did not produce any significant change in the average group difficulty or mean group difference. Maybe our technical sophistication has contributed to the lack of clarity in our results. Some years ago, Lawshe (1969) found that technically more elegant methods were no better than simpler onesin estimating the reliability, the internal consistency, and the weghting of combination scores, and ended the analysis with a warning:

> " the statistical nicety of present methods suggests or implies an order of precision which is not basically inherent in the data; psychological measurements, at this point in time, are quite unreliable; to suggest otherwise by using unwarranted degrees of statistical precision is for the psychologist to delude others and, perhaps, to delude himself (p.122)".

Table 1: Means and Standard Deviations on the Predictor and on the Criterion for Majority and Minority Subjects.

	Majority (N= 68)		Minority (N= 81)		Total (N= 149)	
Test	Mean	S.D.	Mean	S.D.	Mean	S.D.
Criterion	13.35	3.71	12.58	3.14	12.93	3.43
Total Test	24.82	3.03	22.43	3.22	23.52	3.36
Debiased Test	19.21	2.05	18.90	1.88	19.04	1.97
Biased Items	5.62	1.67	3.53	1.90	4.48	2.08

Table 2: Average Difficulty of the Total, the Debiased, and the Biased Tests for Majority and Minority Subjects

Test	Average Difficulty		
	Majority	Minority	Total
Total	82.73%	74.77%	78.40%
Debiased	87.32%	85.90%	86.55%
Biased	70.25%	44.13%	56.20%

Table 3: Correlations between Criterion and Predictor for the Majority, the Minority, and the Total Group of Subjects

Predictor	Validity Coefficients		
	Majority	Mirority	Total
Total Test	.38	.42	.41
Debiased Test	.41	.51	.46
Biased Items	.05	.23	.18
Biased vs Debiased	.33	.48	.39

Table 4: Items detected as most Discrepant by the TID Method among the Items Classified as Biased by the Item Content Criterion.

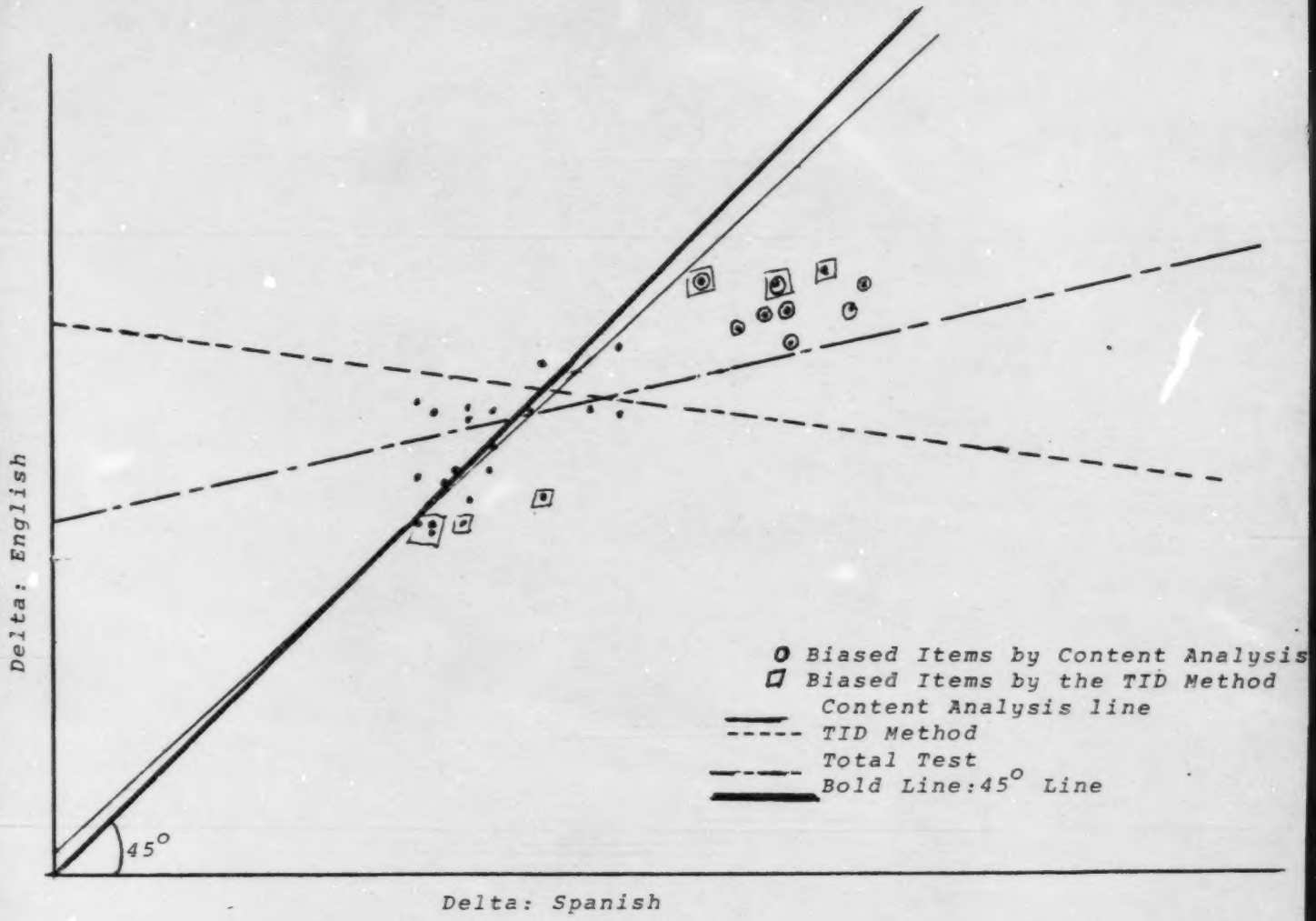
Discrepant Items	Biased Items	Unbiased Items
8 items	2	6
13 items	3	10
18 items	5	13
Itent Content	8	22

Table 5: Level of Difficulty of the Majority and the Minority Means in the Total Test and in the Reduced Test after Removing the Items most Discrepant according to the TID Method.

Discrepant Items	Debiased Test	Difficulty Level		
		Majority	Minority	Difference
0	30	83%	75%	8%
8	22	83%	75%	8%
13	17	83%	76%	7%
18	12	84%	77%	7%
Item Content	22	88%	86%	2%

Table 6: Item of Unacceptable Difficulty among the Items Classified as Biased by the Item Content Criterion

Difficulty Difference	Biased Items(8)	Unbiased Items(22)
10 percentage points	7	4
15 percentage points	7	1
20 percentage points	7	1



Figurel: Graphic Representation of the Delta Values of the 30 Items Vocabulary Test of the CAT for Spanish and English Dominant Students and the Elipse's Major Exes for the 30 Items and the 22 Items Considered Unbiased by the Content-Analysis Criterion and by the TID Method.

BIBLIOGRAPHY

Anastasi, A. Psychological Testing (5th Edition). McMillan Publishing Co.,Inc. New York, 1982.

Angoff, W.H. & Ford, S.F. Item-race Interaction on a Test of Scholastic Aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Angoff, W.H. Use of Difficulty and Discrimination Indices for Detecting Item Bias. In Beck, R.A. (Ed.), Handbook of Methods for Detecting Test Bias. The Johns Hopkins University Press, Baltimore, Maryland, 1982.

California Achievement Test. CTB/McGraw Hill, California, 1977.

Cleary, T.A. Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cole, N.S. Bias in Testing. American Psychologist, 1981, 36, 1067-1077.

Drasgow, F. Biased Test Items and Differential Validity. Psychological Bulletin, 1982, 92, 526-531.

Gordon, E.W. & Terrell, M.D. The Changed Social Context of Testing. American Psychologist, 1981,36, 1167-1171.

Green, D.R. & Draper, J.F. Exploratory Studies of Bias in Achievement Tests. Paper presented at the annual meeting of the American Psychological Association, Honolulu, September 1972 (ERIC Document Reproduction Service, No.070-794).

Humphreys, L.G. Statistical Definitions of Test Validity for Minority Groups. Journal of Applied Psychology, 1973, 58, 1-4.

Humphreys, L.G. & Taber, T. Ability Factors as a Function of Advantaged and Disadvantaged Groups. Journal of Educational Measurement, 1973, 10, 107-115.

Ironson, G.H. Use of Chi-square and Latent Trait Approaches for Detecting Item Bias. In Beck, R.A. (Ed.), Handbook of Methods for Detecting Test Bias. The Johns Hopkins University Press, Baltimore, Maryland, 1982.

Jensen, A.R. Bias in Mental Testing. The Free Press, New York, 1980.

Lawshe, C.H. Statistical Theory and Practice in Applied Psychology. Journal of Personnel Psychology, 1969, 22, 117-124.

Lord, F.M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.

Matarazzo, J.D. Wechsler's Measurement and Appraisal of Adult Intelligence. Baltimore: Williams & Wilkins, 1972.

Nungester, R.J. An Empirical Examination of Three Models of Item Bias. Unpublished Doctoral Dissertation, Florida State University, 1977.

Rudner, L.M. A Closer Look at Latent Trait Parameter Invariance. Educational and Psychological Measurement, 1983, 43, 951-955.

Scheuneman, J.A. A Method of Assessing Bias in Test Items. Journal of Educational Measurement, 1979, 16, 143-152.

Scheuneman, J.A. A Posteriori Analysis of Biased Items. In Beck, R.A. (Ed.), Handbook of Methods for Detecting Test Bias. The Johns Hopkins University Press, Baltimore, Maryland, 1982.

Shepard, L., Camilli, G., & Averill, M. Comparison of Procedures for Detecting Test-item Bias with Both Internal and External Ability Criteria. Journal of Educational Statistics, 1981, 6, 317-375.

Sternberg, R.J. & Powell, J.S. Comprehending Verbal Comprehension. American Psychologist, 1983, 38, 878-893.

Wigdor, A.K. & Garner, W.R. (Eds.). Ability Testing: Uses, Consequences, and Controversies. National Academy Press, Washington, D.C., 1982

Wright, B.D. & Stone, M.H. Best Test Design. Chicago, Mesa Press, 1979.